OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

CDN 边缘节点容器调度的实践

黄励博（huangnauh）

又拍云系统开发高级工程师

又拍云 UPYUN.COM

边缘容器云简介


又拍云
产品
解决方案
价格
帮助与工具
小拍日志
Open Talk
热门专题
登录
注册
控制台
容器云
又拍云容器云是基于 Docker 的分布式计算资源网，节点分散在全国各地及海外，提供电信、联通、移动和多线网络，融合微服务、DevOps 理念，满足精益开发、运维一体化，大幅降低分布式计算资源构建复杂度，大幅降低使用成本。
联系我们
10101
01010
010

Mesos

OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

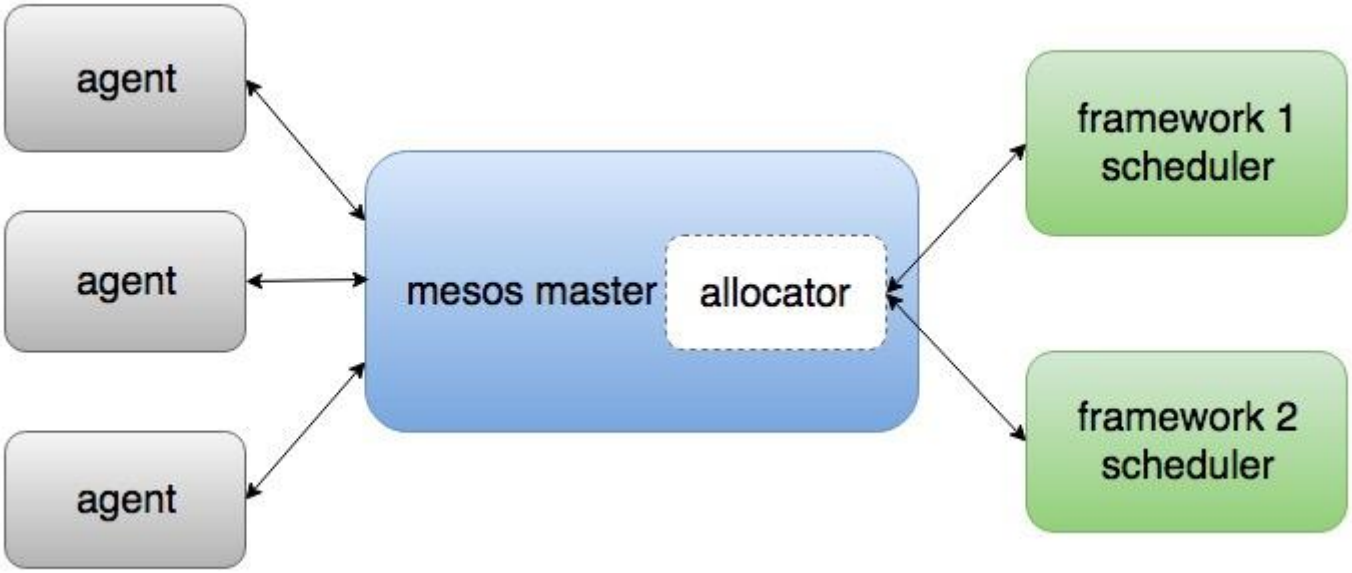
Image credit: mesos.apache.org

- 官方称之为分布式系统内核, 它把数据中心的 CPU、内存、磁盘等抽象成一个资源池

又拍云 UPYUN.COM

Mesos 架构

OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台



- 各个 Agent 启动后, 向 Master 注册，携带统计资源, 由 Master 决定给每个框架多少资源
- 每个框架收到资源后, 根据自身任务需求, 调度任务的资源分配

又拍云 UPYUN.COM

边缘容器调度方案

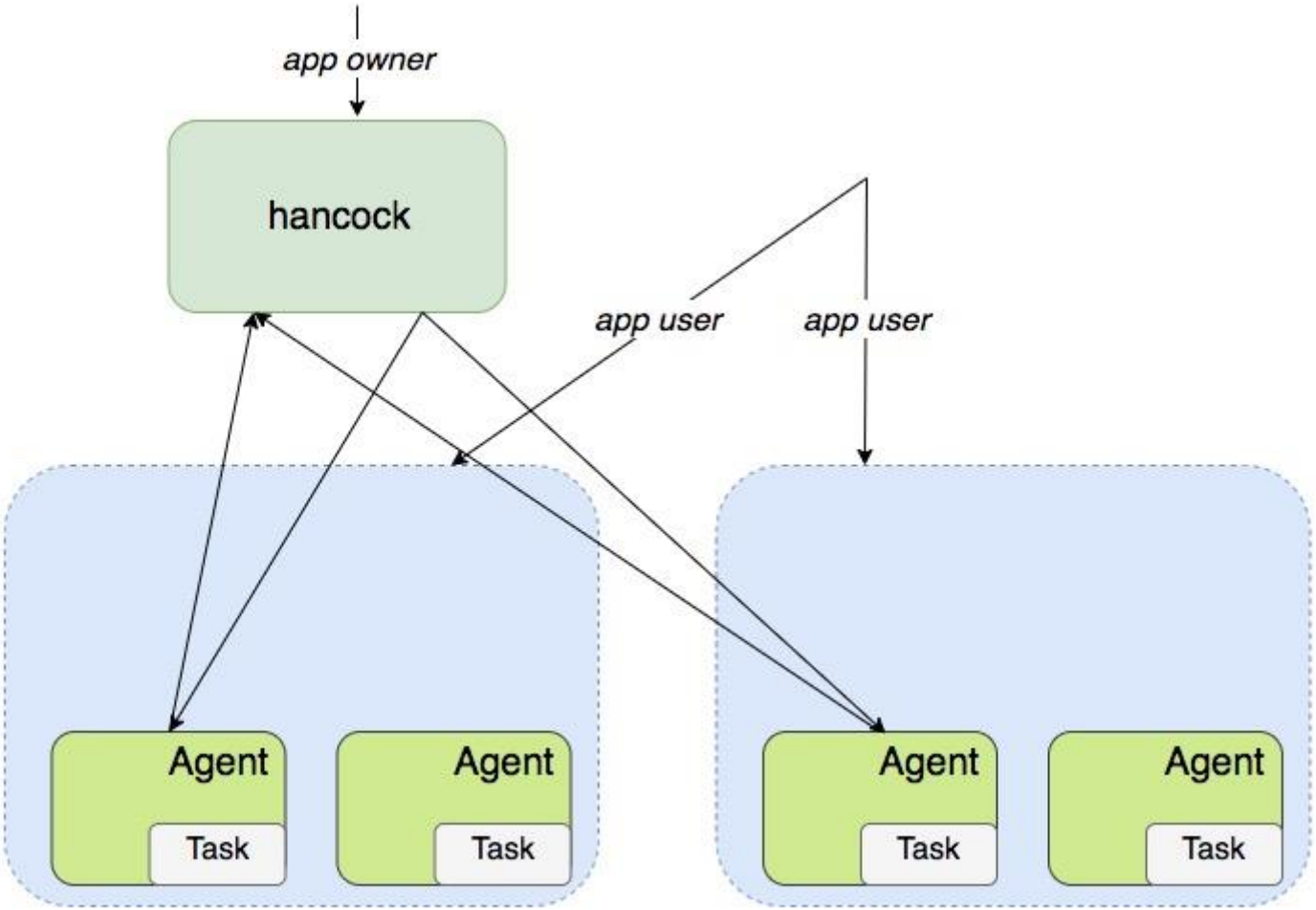
OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

- Master 负责集中调度，统一管理
- Agent 部署在各个边缘节点
- 运行长期服务，支持故障恢复
- 容器网络隔离
- 负载均衡
- 其他一些定制化需求
- ……

又拍云 UPYUN.COM

Master - Agent

OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台



又拍云 UPYUN.COM

消息处理

OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

- 上报消息 Agent -> Master
- 下发指令 Master -> Agent

又拍云 UPYUN.COM

Resource 消息

OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

Agent 启动时上报 Resource 消息， 之后每隔一段时间会上报一次

字段	描述
Node	节点名
Hostname	主机名
Network	网络信息（ip 运营商 带宽等）
Cpu	cpu 大小
Mem	内存大小
Disk	磁盘信息
Port	可用端口(随机分配 + 指定分配)
PortRange	可用端口段(指定分配)

又拍云 UPYUN.COM

Offer 消息

OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

- 每隔一段时间上报一次， 保持 Agent - Master 联系
- Master 会根据资源信息和 offer 的实时信息进行资源调度

字段	描述
Node	节点名
Hostname	主机名
Load	当前 load 信息（用于 Master 调度）
Network	当前网络信息（用于 Master 调度）
CPU	当前可用 Cpu 信息（可选）
Mem	当前可用内存信息（可选）
Disk	当前可用磁盘信息（可选）

又拍云 UPYUN.COM

Update 消息

OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

- 提供实例的状态变更，运行成功， 运行失败等

字段	描述
Node	节点名
Hostname	主机名
TaskId	实例标识
State	实例状态
Message	消息描述
Ip	地址信息

又拍云 UPYUN.COM

Containers 消息

OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

- 提供机器容器列表， 供 Master 检查，防止僵尸任务和遗漏任务

字段	描述
Node	节点名
Hostname	主机名
Containers	容器列表

又拍云 UPYUN.COM

下发指令

OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

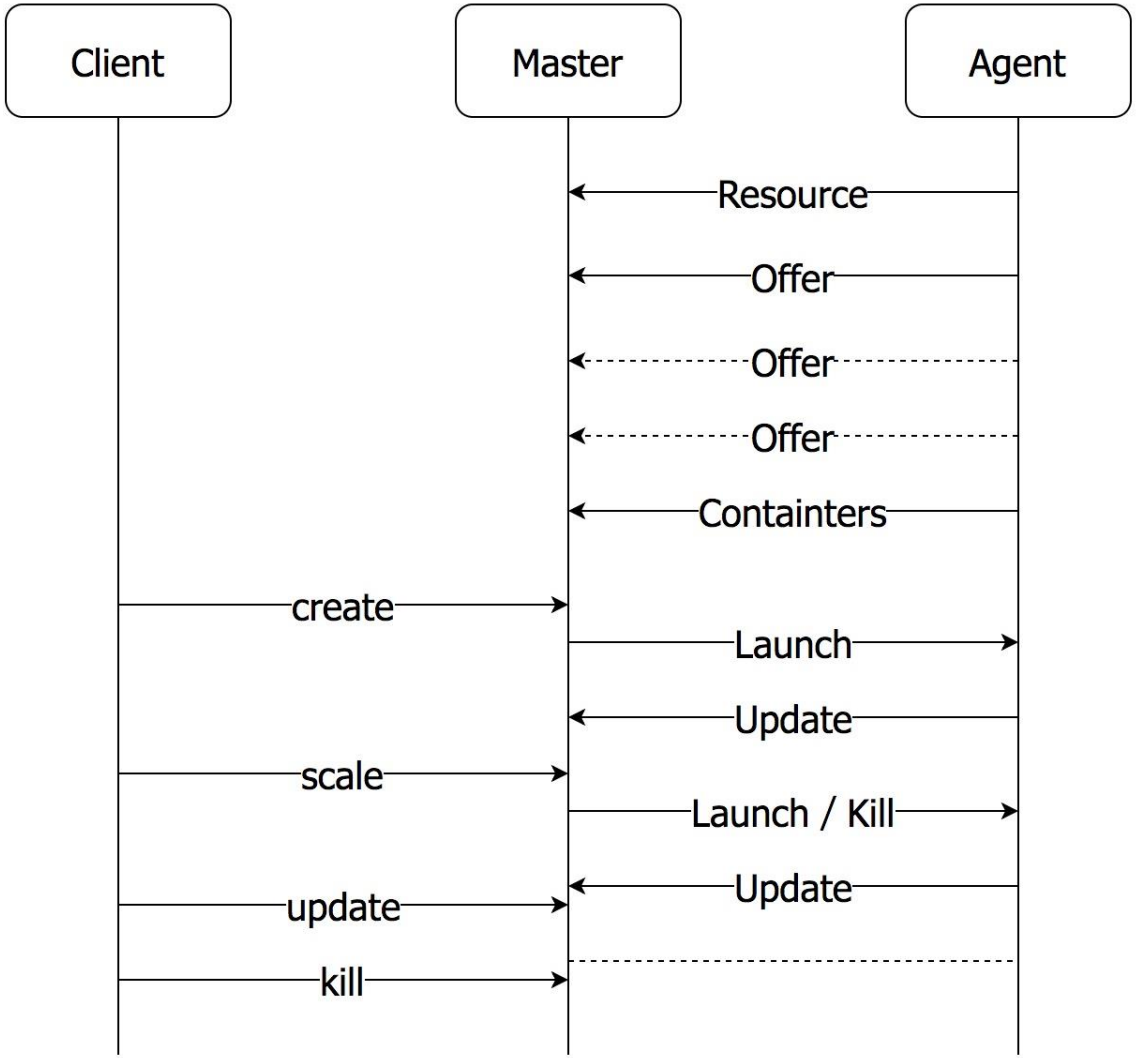
- 实例操作：创建、修改、更新、删除等
- 镜像操作：拉取、查询等

指令	描述
Launch	启动实例
Restart	重启实例
Kill	停止实例
Pull	预拉取镜像
Get	镜像信息

又拍云 UPYUN.COM

处理流程

OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台



又拍云 UPYUN.COM

调度策略

OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

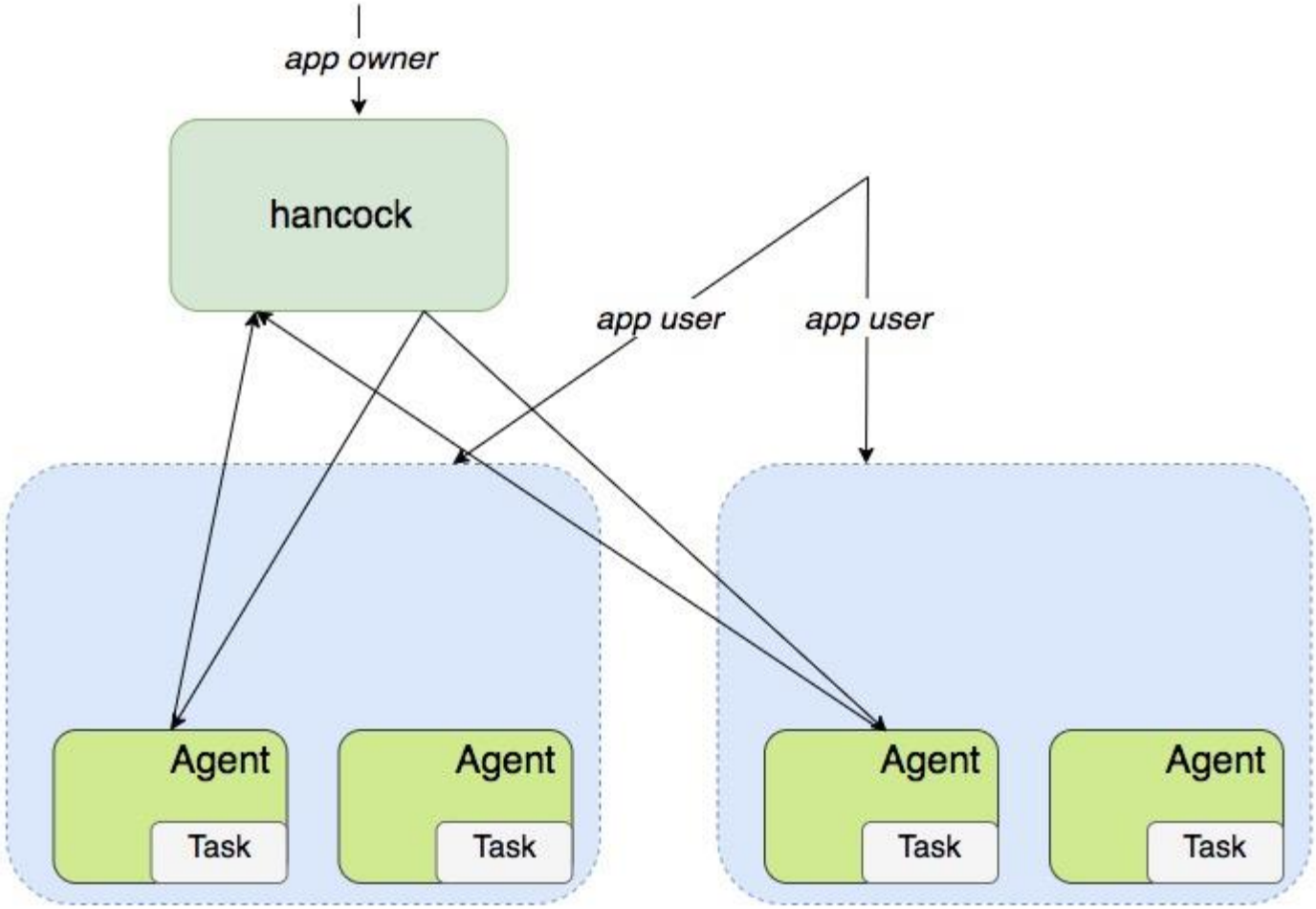
Master 根据 Agent 上报的消息, 可以自定义策略来决定任务调度方案

- 寻找满足条件的（Node Cpu Mem Disk Port 等）机器
- 动态调度: 根据带宽/load 等指标
- 随机调度: 每次调度尽可能让同一服务的各个实例分布到节点的不同机器
- ……

又拍云 UPYUN.COM

版本 1

OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

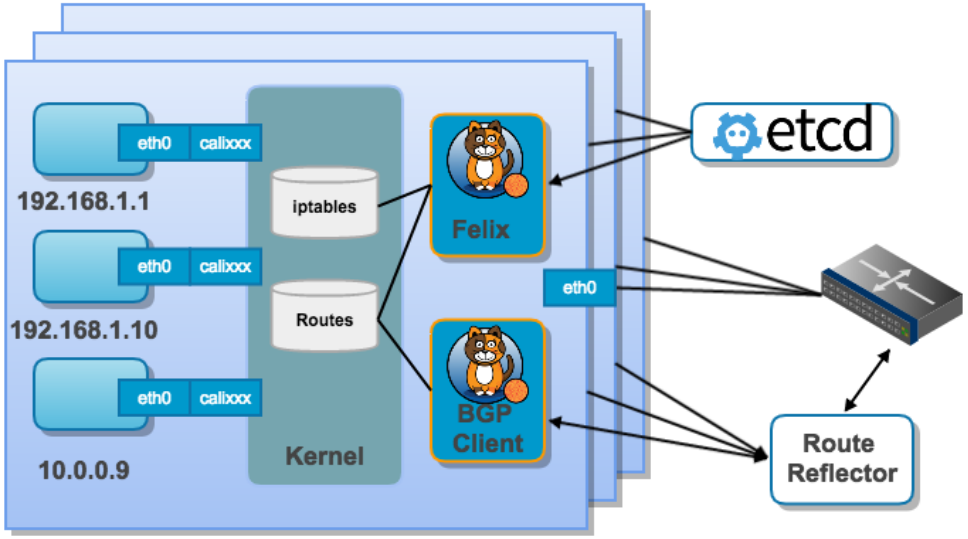


又拍云 UPYUN.COM

容器网络

OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

calico 是一个基于 BGP 路由协议的三层通信模型



- BGP client 每个节点和集群其他节点建立 BGP peer 连接， O(n^2)
- felix 负责更新网络相关配置 Routes iptables
- etcd 分布式的 kv 数据库，保存网络元数据

又拍云 UPYUN.COM

calico

OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

- 三层互联，不需要报文封装
- 访问控制， 满足隔离网络， 隔离容器
- 网络规模限制， iptables 和路由表项限制

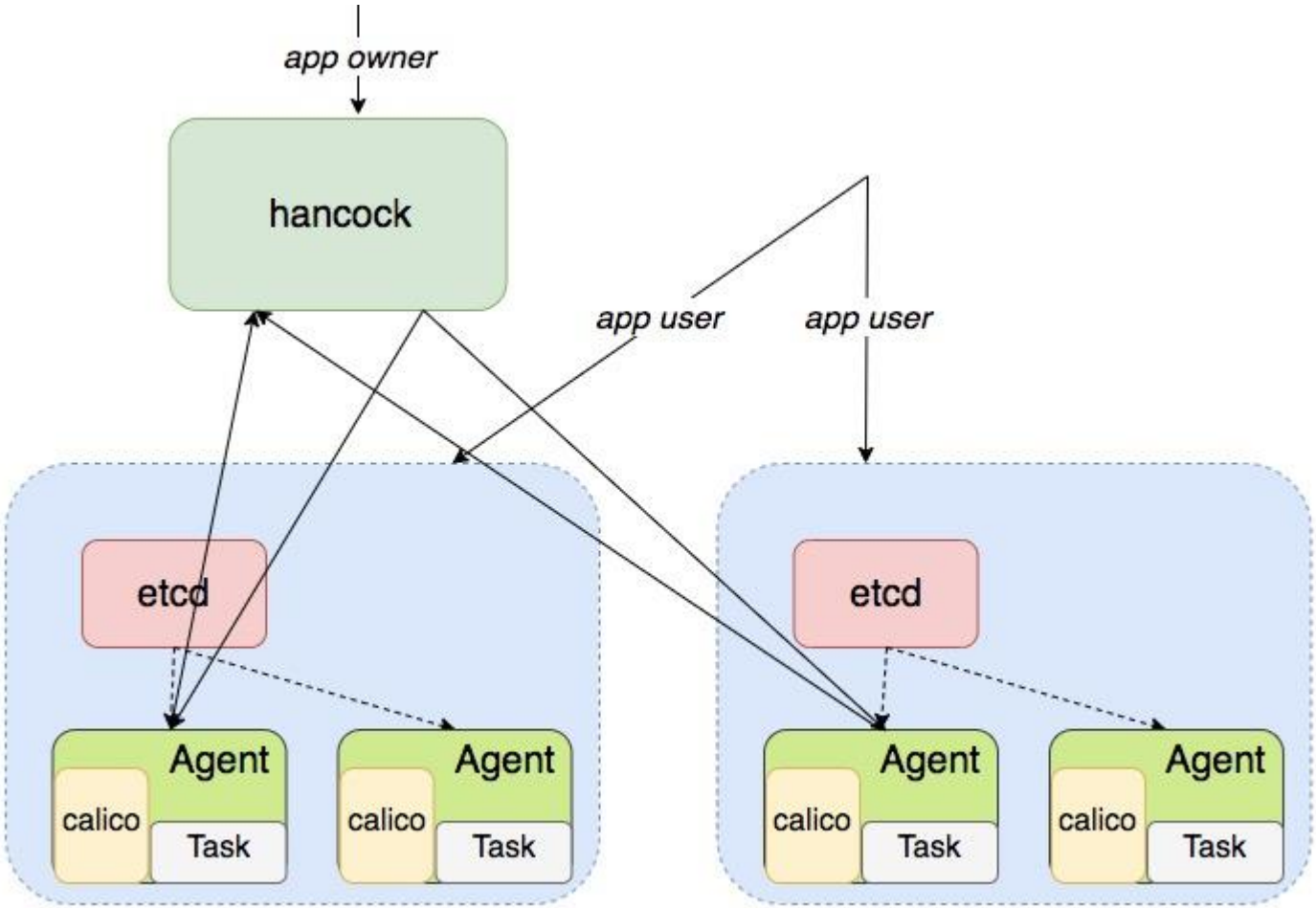
```
[root@DCK-JS-TAZ-010 ~]# calicoctl node status
Calico process is running.

IPv4 BGP status
+---------------+-------------------+-------+----------+-------------+
| PEER ADDRESS  |     PEER TYPE     | STATE |   SINCE  |    INFO     |
+---------------+-------------------+-------+----------+-------------+
| 192.168.18.8  | node-to-node mesh | up    | 2018-01-17 | Established |
| 192.168.18.9  | node-to-node mesh | up    | 2018-01-17 | Established |
| 192.168.18.11 | node-to-node mesh | up    | 2018-01-17 | Established |
+---------------+-------------------+-------+----------+-------------+
```

calico 网络集群

又拍云 UPYUN.COM

版本2

OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台



又拍云 UPYUN.COM

访问入口

- 指定映射端口 Port PortRange
- 基于 **ngx_lua** 的动态负载均衡方案： **Slardar**

Hub, Inc. [US] https://github.com/upyun/slardar

Description

Slardar is a HTTP load balancer based on Nginx and lua-nginx-module, by which you can update your upstream list and run lua scripts without reloading Nginx.

This bundle is maintained by UPYUN(又拍云) Inc.

Because most of the nginx modules are developed by the bundle maintainers, it can ensure that all these modules are played well together.

The bundled software components are copyrighted by the respective copyright holders.

服务动态选择

OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

```
if subsystem == 'http' then
    skey = ngx.var.host
elseif subsystem == 'stream' then
    skey = ngx.var.server_port
end

local peer, err = checkups.select_peer(skey)
```

```
if not peer then
    ngx.log(ngx.ERR, "select peer failed, ", err)
    return
end

local ok, err = balancer.set_current_peer(peer.host, peer.port)
```

又拍云 UPYUN.COM

注册机制

OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

- 负载均衡策略， 默认支持轮询 hash
- 健康检查策略， 默认支持 tcp http mysql 等协议

又拍云 UPYUN.COM

健康检查

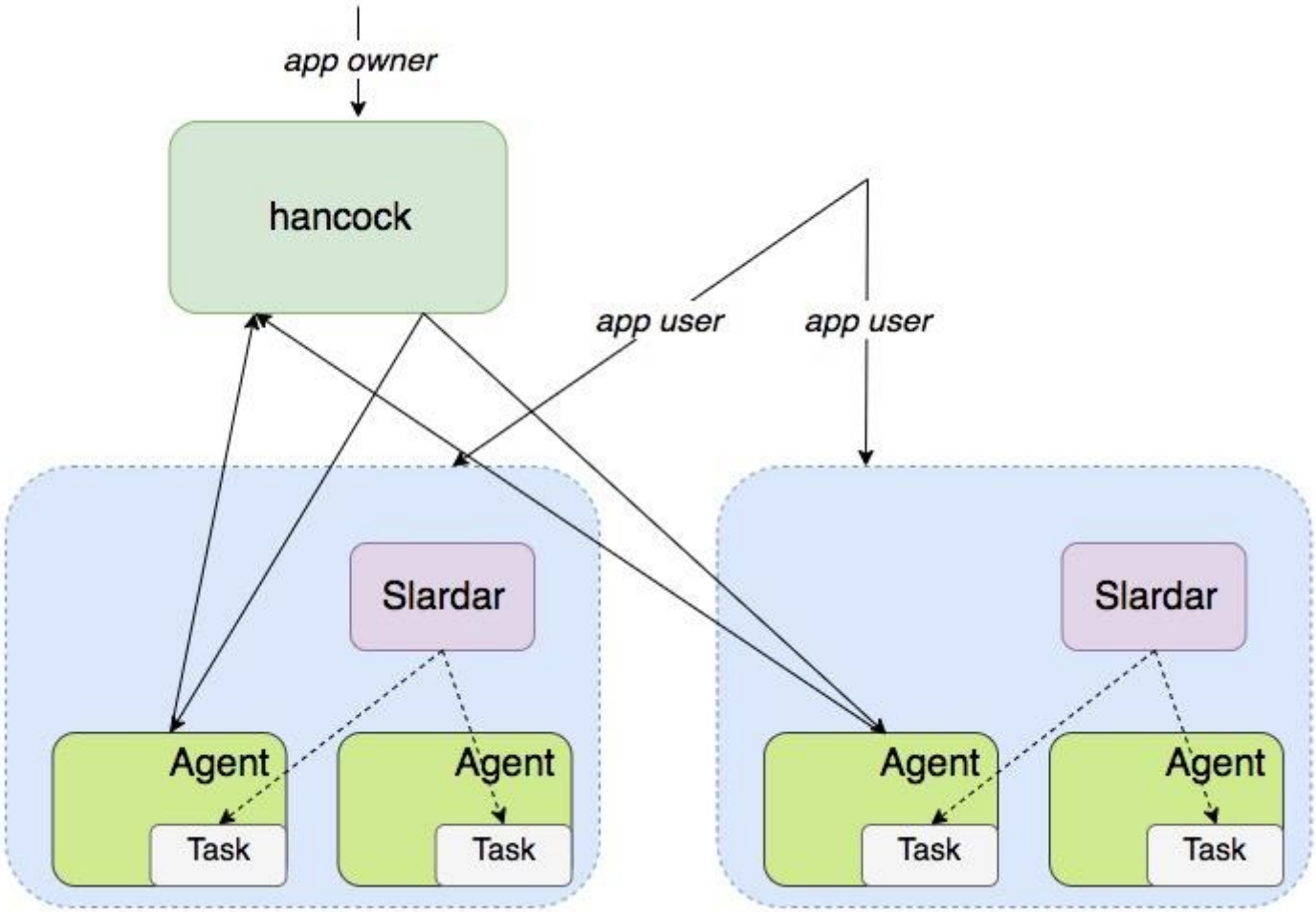
OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

```
{
    -- checkups heartbeat timer is alive.
    "checkup_timer_alive": true,

    -- last heartbeat time
    "last_check_time": "2018-05-17 13:09:40",

    -- status for node-dev cluster
    "cls:node-dev": [
        [
            {
                "server": "node-dev:10.0.5.108:1234",
                "weight": 1,
                "fail_timeout": 30,
                "status": "ok",
                "max_fails": 6
            }
        ]
    ]
}
```

又拍云 UPYUN.COM

版本 3

OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

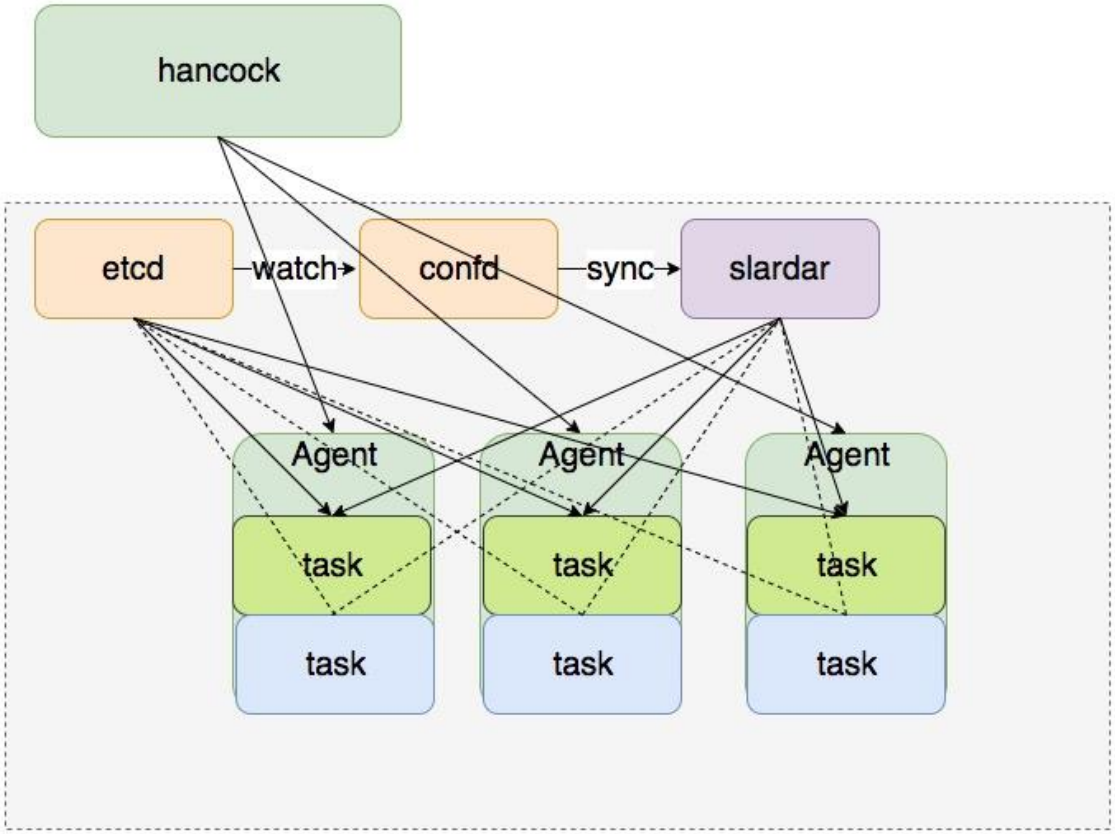


又拍云 UPYUN.COM

零停机更新

OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

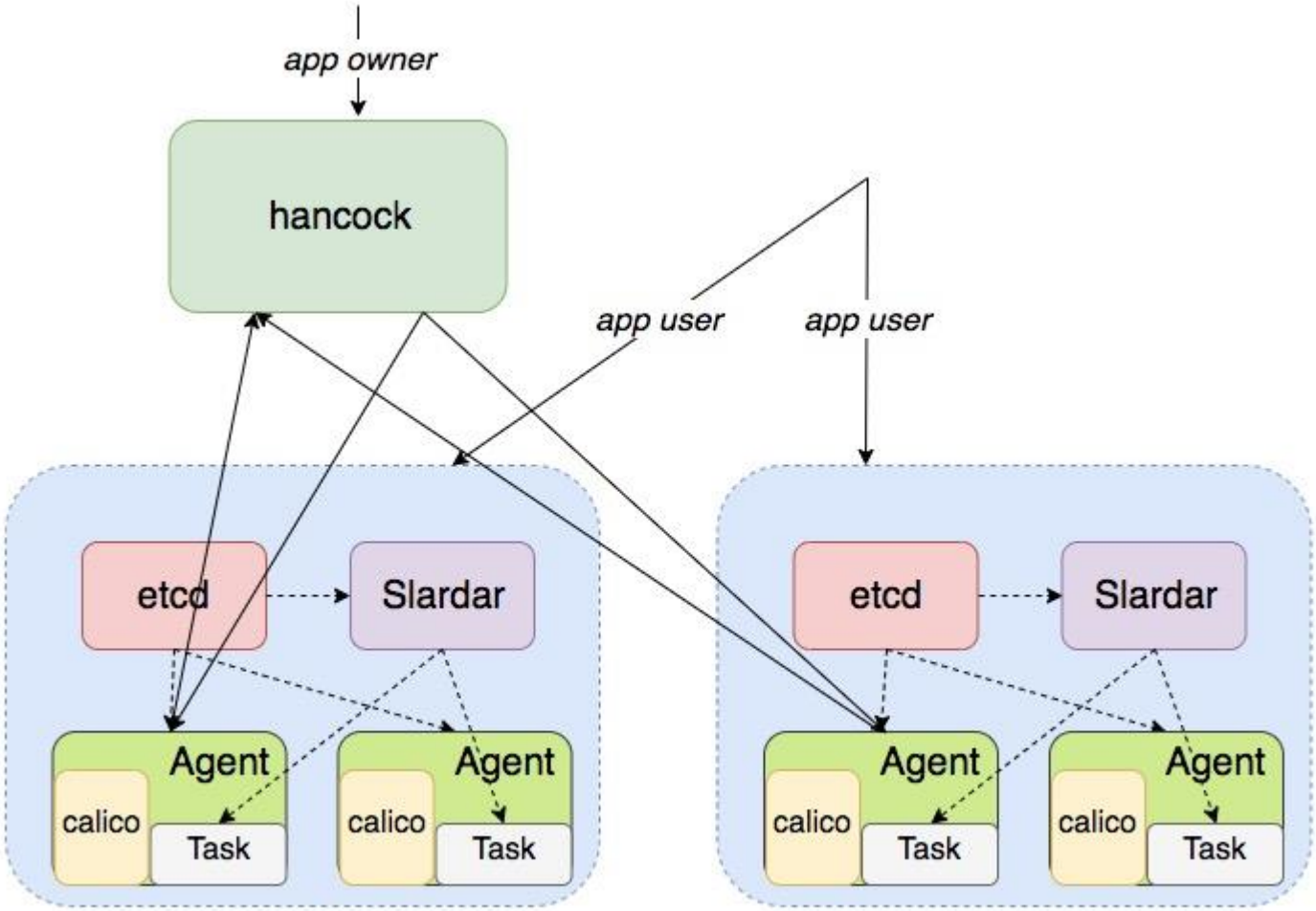
通过允许两个不同版本的服务同时运行来实现蓝绿更新



又拍云 UPYUN.COM

版本 4

OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台



又拍云 UPYUN.COM

Master 高可用

OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

通过 **Raft** 分布式一致性协议实现高可用

[hashicorp/raft]

- 领导选举: 心跳机制来触发选举, term 充当逻辑时钟的作用
- 日志复制: 领导者把一条指令(能被复制状态机执行)附加到日志中，发起附加条目 RPC 请求给其他角色
- 强领导者：日志条目只从 leader 发送给其他的服务器

又拍云 UPYUN.COM

raft 领导选举

OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

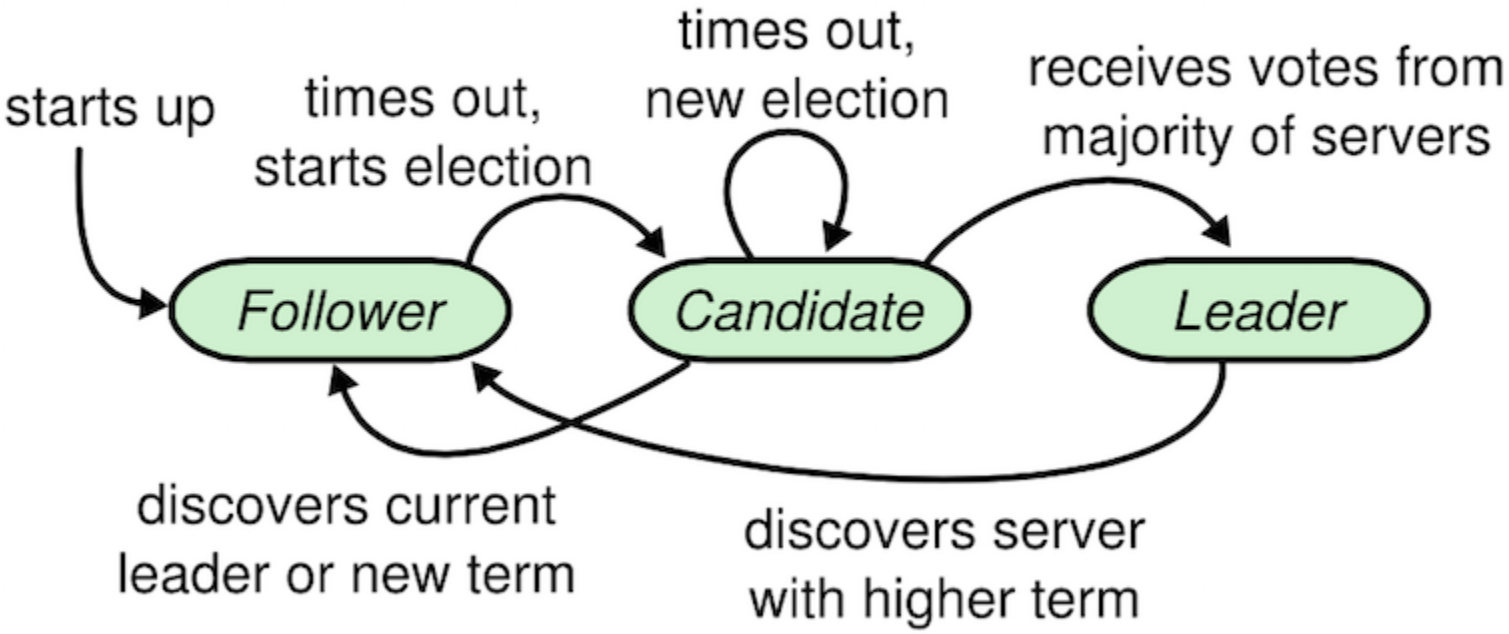


Image credit: [the Raft paper]

又拍云 UPYUN.COM

raft 日志复制

OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

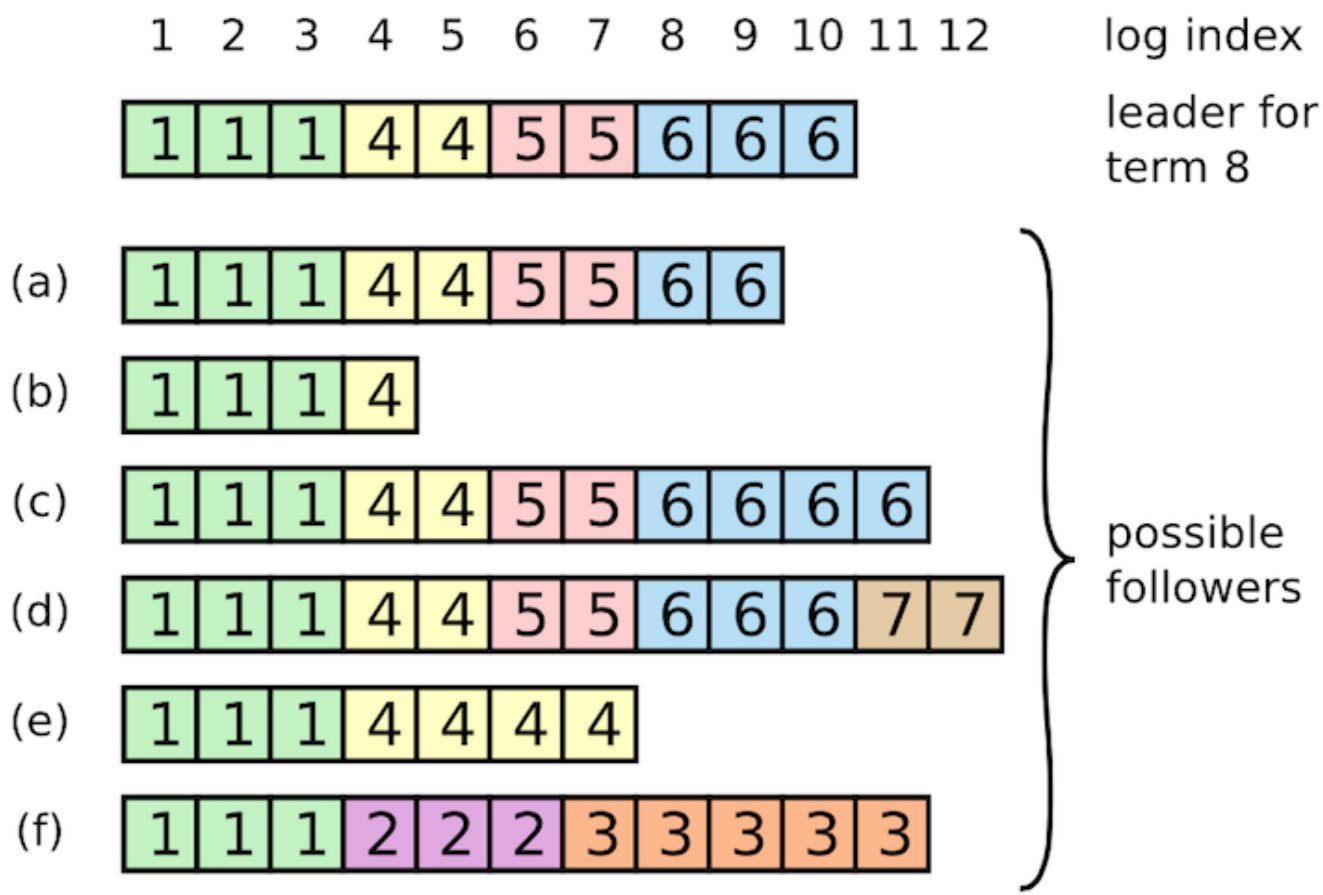
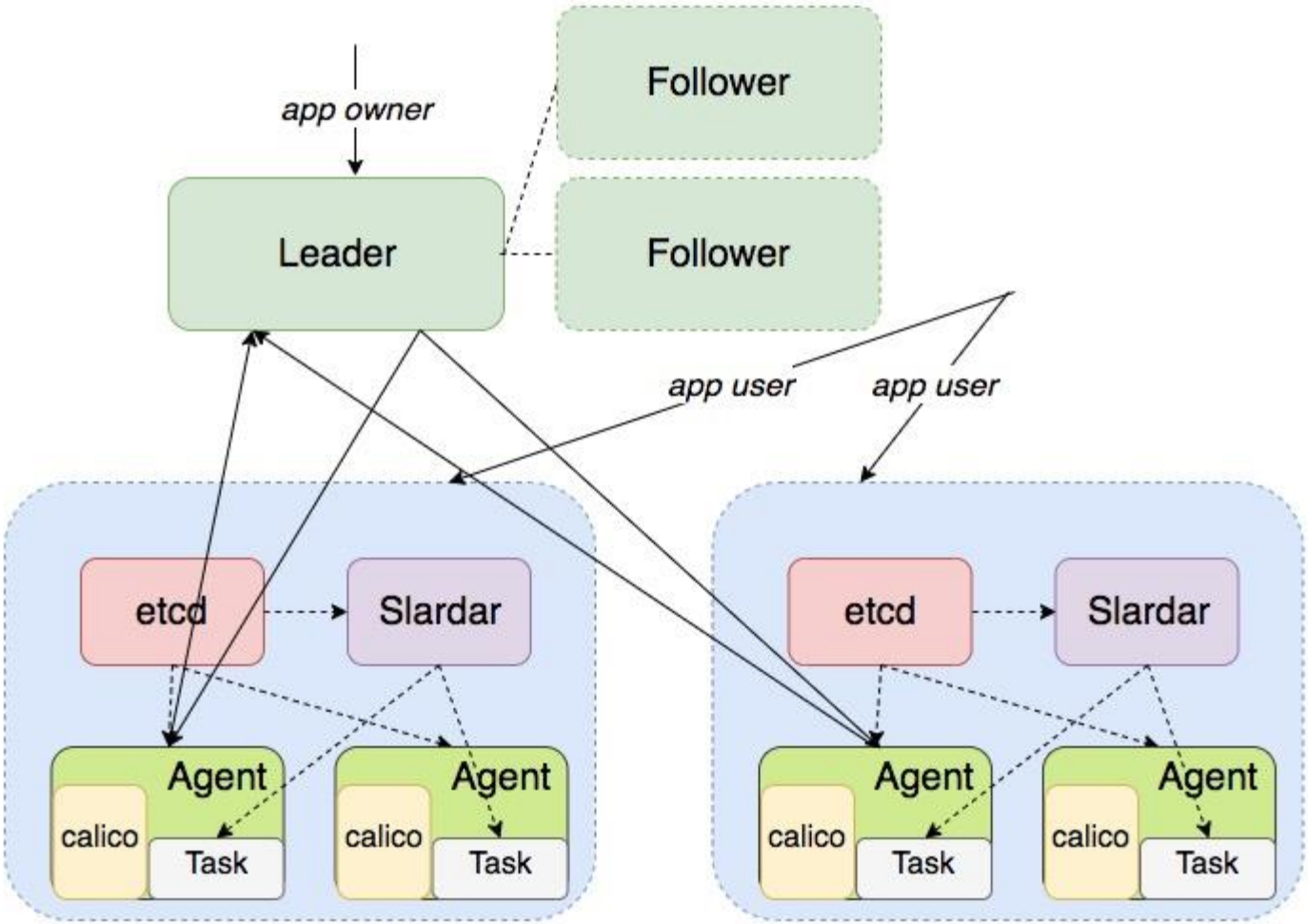


Image credit: [the Raft paper]

又拍云 UPYUN.COM

版本 5

OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台



又拍云 UPYUN.COM

事件通知

OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

服务通过注册回调通知地址，来实时获取各个事件

- 实例状态变更
- 服务状态变更

又拍云 UPYUN.COM

监控

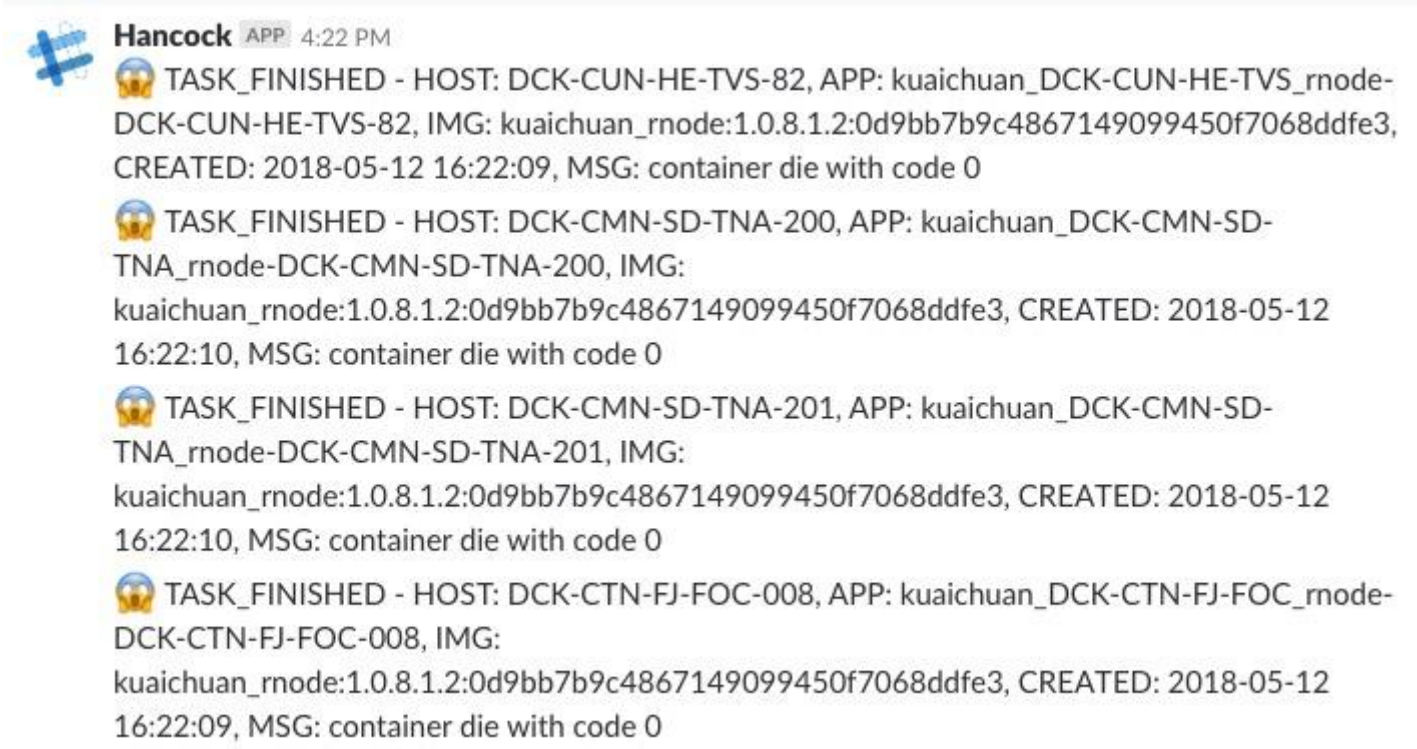
OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

Agent 会收集 Metrics 到 InfluxDB， 由 Grafana 展示

又拍云 UPYUN.COM

告警

- 任务相关：比如 任务失败时, 发送实时通知



告警

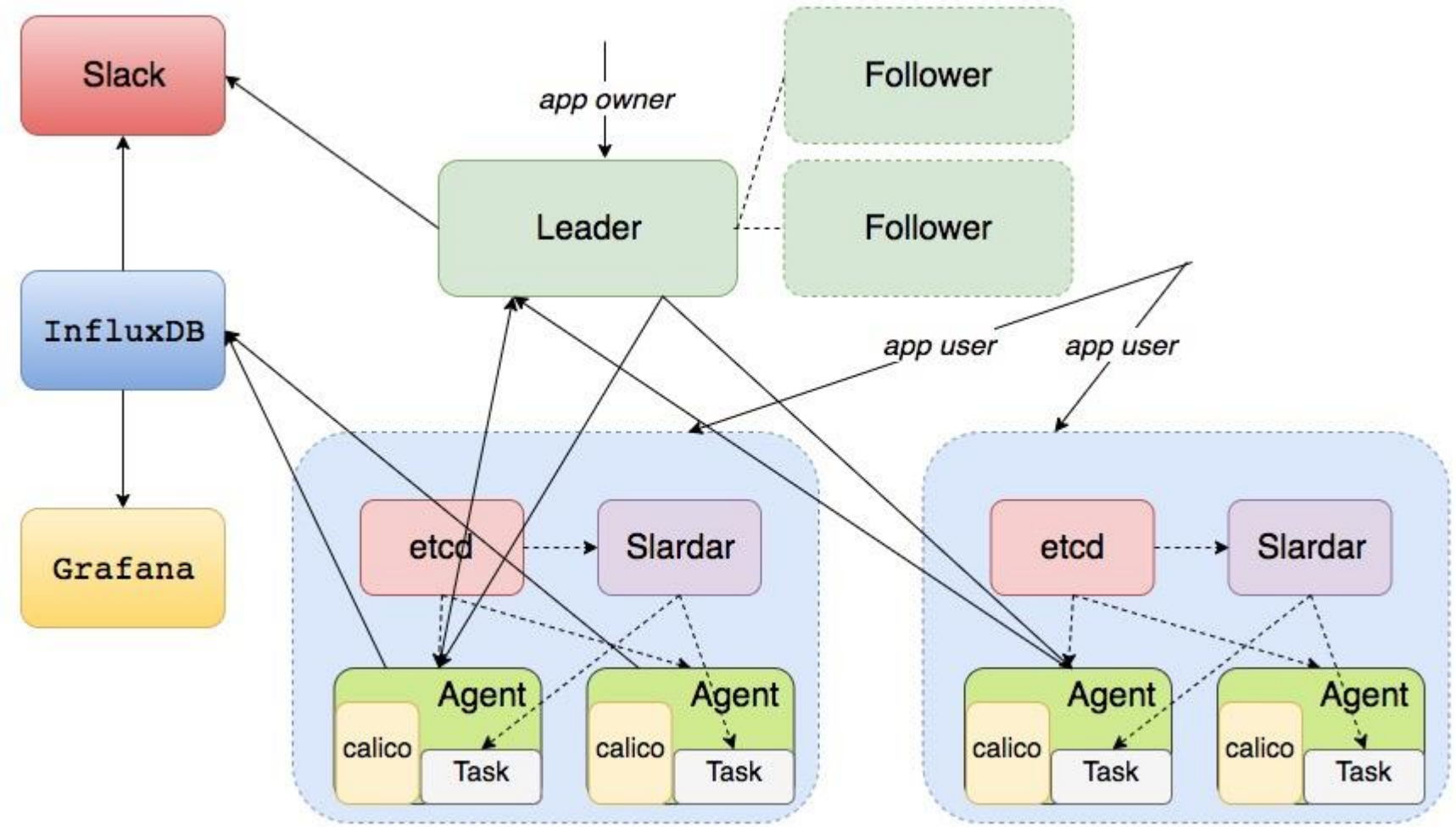
OSC源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

- 监控相关：比如 设定的指标异常时，发送实时通知

又拍云 UPYUN.COM

hancock

OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台



又拍云 UPYUN.COM

OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识共享平台

THANK YOU !

又拍云 UPYUN.COM